Providing Immersive Virtual Experience with First-person Perspective Omnidirectional Movies and Three Dimensional Sound Field

Kazuaki Kondo†, Yasuhiro Mukaigawa††, Yusuke Ikeda†††, Seigo Enomoto†††,
Shiro Ise‡, Satoshi Nakamura†††, and Yasushi Yagi††

†Academic Center for Computing and Media studies, Kyoto University
Yoshida honmachi, Sakyo-ku, Kyoto, Japan
††The Institute of Scientific and Industrial Research, Osaka University
8–1 Mihogaoka, Ibaraki-shi, Osaka, Japan
†††Spoken Language Communication Group, National Institute of Information and Communications Technology
3–5 Hikaridai, Keihanna Science City, Japan
‡Graduate school of engineering, Department of Architecture and architectural engineering, Kyoto University
Kyotodaigaku-katsura, Nishikyo-ku, Kyoto, Japan
kondo@ccm.media.kyoto-u.ac.jp, {mukaigaw,yagi}@am.sanken.osaka-u.ac.jp
{yusuke.ikeda,seigo.enomoto,satoshi.nakamura}@nict.go.jp,
ise@archi.kyoto-u.ac.jp

Abstract. Providing high immersive feeling to audiences has proceeded with growing up of techniques about video and acoustic medias. In our proposal, we record and reproduce omnidirectional movies captured at a perspective of an actor and three dimensional sound field around him, and try to reproduce more impressive feeling. We propose a sequence of techniques to archive it, including a recording equipment, video and acoustic processing, and a presentation system. Effectiveness and demand of our system has been demonstrated by ordinary people through evaluation experiments.

Keywords: First-person Perspective, Omnidirectional Vision, Three Dimensional Sound Reproduction, Boundary Surface Control Principle

1 Introduction

High realistic scene reproduction provides audiences rich virtual experiences that can be used for sensory simulators and multimedia amusements. For example, current cinemas install advanced capturing method, video processing, audio filtering, presentation system to give immersive feeling as they were in the target scene. The most important issue for providing such feeling is to capture and present target scenes as these were. In this paper, we focus on following three functions for that.

Preserving observation perspective: observation perspective can be categorized into third-party perspective and first-person perspective. The former corresponds to objectively capturing a scene, which can effectively conveys structure of the scene and the story line. The latter perspective can be captured by a recording device placed at a character's position, which is good at providing immersive feeling. Examples are to attach a compact video camera to ones head, and actors/actress role as if a video camera was a person.

Preserving wide range(Omnidirectional) visual feeling: A wide range video boosts realistic feeling. A Panoramic video on a wide screen is a typical approach. But it is not always enough because of not considering temporal change and individual difference of audience's observation direction. We focus on capturing and displaying omnidirectional videos in order to adapt to these situations. Although reconstruction of a 3D is also a effective approach, we here treat only an omnidirectional property instead of the combination of them.

Preserving 3D acoustic feeling: Audio reality strongly depends on sounds coming from which directions and how distances. Thus it is important to reproduce 3D sound field including positions of sound sources. Usual approaches are using stereo or 5.1 ch system, but these provide insufficient reproduction for a specific position and direction. We focus on relaxing those listening limitations as watching one discussed in the visual feeling.

Although these functions have been individually attacked in conventional approaches, we do not find any total system that covers all of them. In this paper, we design a special recording device, discuss media processing, and develop a presentation system, in order to satisfy the three functions.

2 Recording System

2.1 Wearable Omnidirectional Camera

We here assume three requirements that a video recording device should have.
- It can capture high resolution and uniform omnidirectional videos.
- Its optical center and a viewpoint of a wearer are at the same position.
- It can be easy to wear and to act for an enough long time.

Unfortunately, conventional approaches to capture outdoor scenes as omnidirectional videos [5, 7] do not satisfy all of the above requirements. These did not consider to capture a scene from a character's viewpoint, and approximate the viewpoint matching with omnidirectional cameras mounted on the head. Furthermore, needs of an additional equipment for recording and power supply prevent the third requirements. A wearable omnidirectional camera has been proposed [9] for life-log recording. But it also has the viewpoint mismatch, and additionally low resolution problem. For these reason, we had proposed a special wearable camera system named FIPPO [10]. FIPPO is constructed by four optical units consisting of a handy type video camera and curved and flat mirrors (Fig. 1(a)). It captures omnidirectional videos from first-person perspective without any additional equipments and wired supply. Following descriptions briefly explain design of a single optical unit in FIPPO.

We start at an objective projection defined by correspondences between pixels on the image plane and rays running in the scene. Considering uniform resolution of the panoramic scene whose FOV are $[\theta_{min}, \theta_{max}]$ along azimuth angle and $[tan\phi_{min}, tan\phi_{max}]$ along elevation, respectively, the objective projection is formulated as

$$\mathbf{V_s}(u,v) = \begin{bmatrix} tan(\frac{u}{U}(\theta_{max} - \theta_{min}) + \theta_{min}) \\ \frac{v}{V}(tan\phi_{max} - tan\phi_{min}) + tan\phi_{min} \\ 1 \end{bmatrix} \tag{1}$$

in the world coordinate system. (u, v) is the position of a pixel on the image plane whose size is $U \times V$. It determines also a corresponding camera projection $V_c(u, v) = [u, v, -f]^t$ with its focal length f. They, V_s and V_c, should relate a target curved mirror to its reflection ; An objective normal vector field $\mathbf{N_d}(u, v)$ bisects the angle consisting of $\mathbf{V_s}$ and $\mathbf{V_c}$. It is obtained by

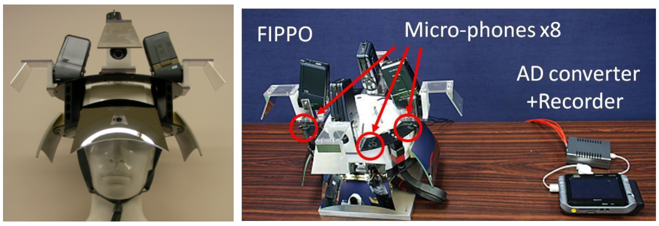
$$\mathbf{N_d} = \mathcal{N}\left[\frac{1}{2}\left(\mathcal{N}[\mathbf{V_s}] + \mathcal{N}[\mathbf{R}\mathbf{V_c}]\right)\right]. \tag{2}$$

with a vector normalizing operator $\mathcal{N}[\mathbf{x}] = \frac{\mathbf{x}}{||\mathbf{x}||}$, and external parameters of the camera $\mathbf{P} = [\mathbf{R} \quad \mathbf{t}]$. The mirror shape is formed so that its normal field is equal to $\mathbf{N_d}$. We use the linear algorithm [6], which equalizes the mirror shape $\mathbf{S}(u, v)$ as the cross products of four-degree spline curves. $\mathbf{S}(u, v)$ is formulated as

$$\mathbf{S}(u,v) = \mathbf{R}\mathbf{V_c}(u,v) \sum_{i,j} C_{ij} f_i(u) g_j(v) + \mathbf{t} \tag{3}$$

where C_{ij} and $f_i(u), g_j(v)$ are control points on the spline curves and four-degree spline bases, respectively. We obtain an optimal shape by solving the linear equations about C_{ij} that are stacks of $\frac{\partial \mathbf{S}}{\partial u} \cdot \mathbf{N_d} = \frac{\partial \mathbf{S}}{\partial v} \cdot \mathbf{N_d} = 0$ because the optimal shape should be perpendicular to the desired normal vector field $\mathbf{N_d}$. Although this algorithm certainly minimizes errors on the normal vector field, it tends to form a bumpy surface. So we apply a smoothing procedure to the shape formed by this algorithm.

The obtained mirror approximates the objective projection. Thus, we check the degree of the approximation. It is evaluated by sufficiency : how much does it cover the required FOV, redundancy : how much does it cover outside the FOV, and uniformity : how uniformly does it distribute the image. If the approximation is sufficient, the design advances to the next step. If not, we adjust the camera parameters to reduce projection errors and return to Eq. (2). Aberrations of the designed optics need to be also checked because the mirror design algorithm does not consider image focusing. The amount of aberration can be estimated with a spot diagram, which is a spread image of a target object on the image plane. We can construct spot diagrams by tracing the rays that go through an aperture of the lens unit. If the aberrations appear to prevent image focusing, we adjust the camera parameters to reduce the aberration and return to the first step. The design process continues until the aberrations are acceptably small.



(a) FIPPO (b) Microphone array and a recorder

Fig. 1. Overview of the recording system

2.2 Microphone array

Methods which have been used for recording and reproducing first-person perspective sound field include head and torso simulator(HATS) and recording with microphones worn on listener's ears. However, in these methods, it is impossible to freely move a listener's head, because the sound signal is reproduced at only two points around the ears. In this paper, we used a sound reproduction system based on boundary surface control(BoSC) principle so that it gives the listener an experience of the sound field from first person's perspective with omnidirectional movies.

Original BoSC system[8] has 70ch microphone array. It is difficult to make microphone array-aided recording of the first-person perspective sound field accompanied by free body movements. Therefore, we simplified the system through reducing the number of channels. The recording system has eight omnidirectional microphones which are installed horizontally around the head of a wearer. It is recommended that the height of microphone is on the same level with the person's ears. The microphones are installed slightly over the top of the head in order to keep the microphones away from the mirrors of FIPPO (Fig. 1(b)). The system is small enough and allows the person to freely move while wearing it. One of the factors contributing to the small system is that the signal is recorded in a handheld PC through a small Bus-Powered USB A/D converter.

3 Media Processing for Making Contents Movie

3.1 Image Processing for Omnidirectional Panorama

Correcting Image Warping Images captured by FIPPO still have some geometric warps, despite of uniform projection being configured as objective one. Calibrations of the distorted projections produced by the entire optical system, including the curved mirrors, allow the images to be corrected. The calibrations were conducted with a particular scene construction in order to homologize image pixels and rays in the world. FIPPO placed at the front of a wide flat panel monitor captures coded patterns that give correspondences between each pixel on the image plane and each 2D position on the monitor. Measurements for

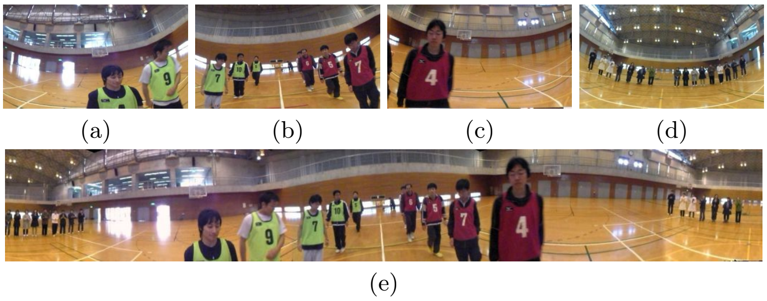


Fig. 2. A result of image unwarping (panoramic). (a)-(b) Input images for each direction, left, front, right, and back (e) Unwarped and mosaiced image.

planes at several depths are necessary for pixel-ray correspondence. When measurements are taken at two depths whose distances d are given, the pixel-ray correspondences can be formulated by

$$\mathbf{ray}(u,v) = \begin{bmatrix} \mathbf{p_1} - \mathbf{p_2} \\ d \end{bmatrix} = \begin{bmatrix} x_1(u,v) - x_2(u,v) \\ y_1(u,v) - y_2(u,v) \\ d \end{bmatrix} \tag{4}$$

where $\mathbf{ray}(u,v) = [r_x, r_y, r_z]^t$, and $\mathbf{p_i} = [x_i, y_i]^t$ denote a ray in the world corresponding to a point(u,v) on the image plane, and a 2D position on the monitor plane, respectively. Figure 2 shows an example of correcting image warping based on the calibration results. Eq. (4) says that directions of rays are determined, but not their position. Thus note that the calibration does not work well for near scene because FIPPO is designed to be approximated as a single viewpoint optical system.

Correcting Color Space It is also necessary to correct chromatic differences that are mainly attributable to individual differences in the cameras. We solved this problem by transforming color spaces under the assumption of an affine transformations between them. It enough approximates relationship of color spaces produced by the same model cameras used in FIPPO. The affine transform is related by

$$\begin{bmatrix} R_m \\ G_m \\ B_m \end{bmatrix} = \begin{bmatrix} p_{11}\, p_{12}\, p_{13}\, p_{14} \\ p_{21}\, p_{22}\, p_{23}\, p_{24} \\ p_{31}\, p_{32}\, p_{33}\, p_{34} \end{bmatrix} \begin{bmatrix} R_n \\ G_n \\ B_n \\ 1 \end{bmatrix} \tag{5}$$

where $[R_k, G_k, B_k]^t$ and p_{ij} are RGB colors of the same object on the k-th camera and coefficients of the affine transformation, respectively. Since Eq. (5) forms three linear equations for one color correspondence, at least four color correspondences are necessary to determine twelve unknowns in p_{ij}. Figure 3 shows the results of chromatic correction. The images in the figure show a neighborhood

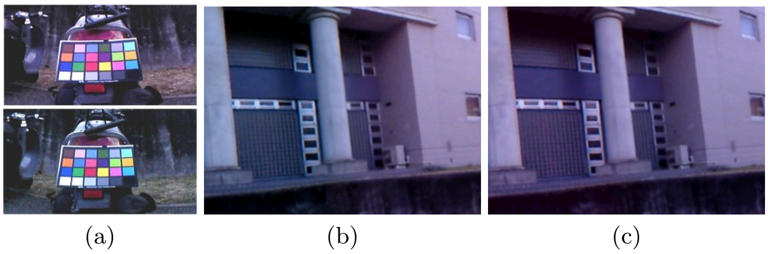


Fig. 3. Chromatic correction. (a) Color checkers captured by different cameras. (b) Mosaiced images without the correction. (c) That with the correction.

of image mosaics. The vertical line at the horizontal center corresponds to the border between two contiguous images. Blue components that were relatively strong assume natural coloring after the correction.

3.2 Reconstructing 3D sound field

Boundary surface control principle It follows from Kirchhoff-Helmholtz integral equation that a control of sound pressures and sound pressure gradients on a boundary of region means a control of sound pressures inside the boundary. Boundary surface control principle removes the problem of ideal sound sources and the restriction of free sound field using Kirchhoff-Helmholtz integral equation and multi-channel inverse system[2]. When applied in the 3D sound reproduction system, the microphones are set at arbitrarily-chosen points within the 3D sound field, and by reproducing the sound pressures recorded at those points in a different location, it becomes possible to accurately reproduce sound field of the area enclosed with the microphones. Therefore, it is different from common transaural system and binaural system. In BoSC system, a listener freely moves his body listening to the sound field which is consistent with the original sound field.

Design method of inverse system Here loudspeakers controlling sound pressures and points which are targets of sound pressure control are referred to as "secondary sound sources" and "control points" respectively. The number of secondary sources and control points are denoted by M and N respectively. Frequency transfer characteristic between ith sound source and jth control point is denoted by $G_{ji}(\omega)$. Recorded signal at primary sound field, output signal from sound source and measured signal at control points are denoted by $X_j(\omega)$, $Y_i(\omega)$ and $Z_j(\omega)$ respectively. The relationship between inputs and outputs of sound reproduction system is as follows.

$$\mathbf{Z}(\omega) = [\mathbf{G}(\omega)]\mathbf{Y}(\omega) = [\mathbf{G}(\omega)][\mathbf{H}(\omega)]\mathbf{X}(\omega) \tag{6}$$

where, $\mathbf{X}(\omega) = [X_1(\omega), \cdots, X_N(\omega)]^T$, $\mathbf{Y}(\omega) = [Y_1(\omega), \cdots, Y_M(\omega)]^T$, $\mathbf{Z}(\omega) = [Z_1(\omega), \cdots, Z_N(\omega)]^T$,

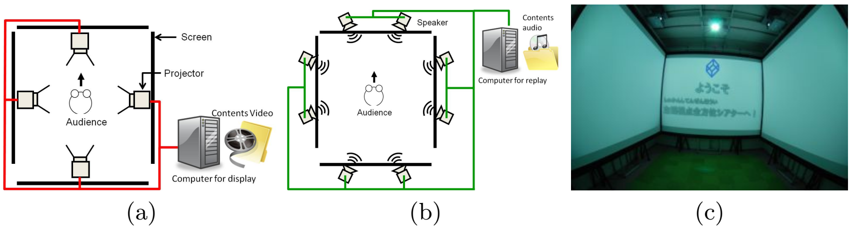


Fig. 4. Omnidirectional theater. (a) Omnidirectional video display system. (b) 3D sound reproduction system. (c) Inner of the theater.

$$[\mathbf{G}(\omega)] = \begin{bmatrix} G_{11}(\omega) & \cdots & G_{1M}(\omega) \\ \vdots & \ddots & \vdots \\ G_{N1}(\omega) & \cdots & G_{NM}(\omega) \end{bmatrix} \mathit{and}\ [\mathbf{H}(\omega)] = \begin{bmatrix} H_{11}(\omega) & \cdots & H_{1N}(\omega) \\ \vdots & \ddots & \vdots \\ H_{M1}(\omega) & \cdots & H_{MN}(\omega) \end{bmatrix}.$$

The purpose of designing inverse filter in the sound reproduction system is to find the inverse filter $[\mathbf{H}(\omega)]$ of $[\mathbf{G}(\omega)]$. When a small error included in $X(\omega)$ and variation of system transfer function$[\mathbf{G}]$ largely effect the value of $\mathbf{Z}(\omega)$, inverse filter $[\mathbf{H}(\omega)]$ becomes unstable. We, thus, designed inverse filter using a regularization which can continuously change the parameter to ease the instability.

4 Presentation System: Omnidirectional Theater

Omnidirectional movies should be displayed all around on viewers with a wide FOV in order to provide immersive feeling. Researchers have proposed omnidirectional display systems for such a situation. These are categorized into personal use equipment[3] and dome or room type systems for multiple persons[1]. We developed the latter type omnidirectional theater to emphasize that multiple audiences share the same feeling. The theater consists of four projectors and four $3m \times 2m$ flat screens standing like walls of a square room(Fig. 4).

Eight loudspeakers surrounding a listener reproduce a recorded sound field based on BoSC principle. The two loudspeakers are set behind each screen which is perforated for a sound. We measured the impulse responses between each loudspeaker and the microphone array which is set inside the theater and has the same alignment with the microphone array which was used for recording. We calculated the inverse filter which has 4096 points length. In order to simplify the calculation of the inverse system, acoustic panels and carpets are installed on the ceiling and the floor of the theater respectively. The sound field inside the microphone array is reproduced by the loudspeakers which have the convoluted signal of calculated inverse filter and recorded signal as the output signal. It is expected that the sound field of a larger region is reproduced so that it is the same as the original sound field[4]. It is also expected that the sound field is more accurately reproduced because the inverse filter compensates not only an

Table 1. Contents of the questionnaire.

Questions about realistic sensation(five-grade scales).
A. Did you got immersive feeling as you were in the scene ? 1. Not at all 2. Not much 3. As usual 4. Fairly 5. Much
B. How much did you feel reality compared with a single front movie ? 1. Not at all 2. Not much 3. As usual 4. Fairly 5. Much
C. How was the image quality ? 1. Bad 2. Not good 3. Normal 4. Good 5. Great
D. How much did you feel reality compared with a stereo sound ? 1. Not at all 2. Not much 3. As usual 4. Fairly 5. Much
E. How was the audio quality ? 1. Bad 2. Not good 3. Normal 4. Good 5. Great
Questions with free-form spaces.
F. What additional features are necessary for the current system ?

attenuation caused by the sound screen but also the acoustic characteristics of a theater.

5 Experiment

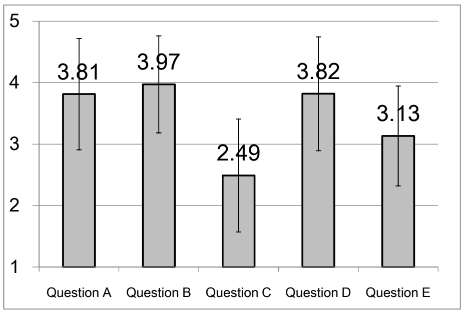
5.1 Configurations

We validated our scene reproduction proposal from viewpoint of immersive feeling through a virtual experiment. Presented first-person perspective contents were (1) daily scenes in the park including fall foliage, water currents, and other persons, and (2) basketball game scenes like shown in Fig. 2.

The experiment had been executed for each group consisting of 3-5 persons in an outreach event held at the National Museum of Emerging Science and Innovation in Tokyo. They experienced the 3 minute video content consisting of the scenes addressed in the above. After that, we conducted questionnaires whose items are listed in Table 1. These are five-grade scale questions and questions with free-form answer spaces. The former is related to realistic sensation that the viewers felt. The latter is to obtain objective opinions about demands of first-person perspective omnidirectional movies and issues that should be improved. We got about 750 valid responses from more than $1,100$ subjects who experienced our system for three days. Since the subjects were in a wide age range, and included groups such as couples, friends, and families, we can expected general and objective evaluations.

5.2 Results and Discussions

We can see that most subjects felt highly realistic sensations from the result shown in Fig. 5(a), which demonstrates the effectiveness of first-person perspective omnidirectional movies. Unfortunately, image quality got a low score. One



(a)

- Quality of the medias
  - Improve image quality
  - (resolution and contrast)
- Video quaking
  - Little sick, nauseous, dizzy
  - Film without head swinging
- Reproduction of spatial sense
  - Desire sense of correct depth
  - Recognize shapes of objects and scenes
- Construction of the screen
  - A cylindrical, a doom, and eight flat screens
  - Additional overhead screen

(b)

Fig. 5. (a) Results of the five-grade scale questions. 1 and 5 denote lowest and highest scores, respectively. (b) Representative answers written in the free-form spaces.

reason is optical construction of FIPPO. Since rays from a scene are reflected multiple times to be projected on image planes, light quantity decreases in each reflection, resulting in low quality images. The mirrors used in the prototype FIPPO are covered with low reflective material. This problem can be solved by using high reflective material. The other reason is less image contrast at the display stage. There are some causes such as output contrast of the projectors and inter reflections between the screens.

The representative opinions written in the free-form space are listed in Fig. 5(b). Video quaking that means image shakes and blurs caused by rapid ego-motions were pointed out as problems that should be solved. Some subjects said that they felt nauseous or got dizzy. In a way, our system can truly reproduce a first-person perspective, including head swing, but this is actually worse to be displayed to static viewers. Recording a head state with a gyro sensor or ego-motion estimation algorithms that use horizontal cyclic property of omnidirectional movies will help with video stabilization. Some subjects said that a cylindrical screen should be used instead of the four flat screens that gives an incorrect sense on depth. A fundamental approach is needed to provide a spatial sense not considered in the proposed method. Omnidirectional scenes must be spatially constructed, which requires special capturing equipment. A three dimensional display all around viewers is also a challenging issue.

6 Conclusion

In this paper, we proposed a virtual experience system that provides high realistic feeling to audiences with omnidirectional videos and 3D sounds captured from a first-person perspective. Contents data are captured by a specially designed wearable equipment consisting of catadioptric imaging systems and a microphone array. Audio and visual media processing for providing high realistic feeling and a presentation system are also discussed. Its performance have been evaluated by experiences of more than 1,000 ordinary persons. At the same time of expected results on high realistic and immersive feeling, we got several problems such as video quaking, media quality, and sense of depth given by video. These problems are now being attacked by other proposals. Thus the combination with them will give more attractive virtual experiences.

Acknowledgment

This work is supported by the Special Coordination Funds for Promoting Science and Technology of Ministry of Education, Culture, Sports, Science and Technology. The authors wish to thank SANYO Electric Co. Ltd. for providing the specially modified portable video cameras.

References

1. C. Cruz-Neria, D. J. Sandin, and T. A. DeFanti, “Surround-Screen Projectorion-Based Virtual Reality : The Design and Implementation of the CAVE”, Proc. of Int. Conf. on Computer Graphics and Interactive Techniques(SIGGRAPH1993) 1993, pp. 135–142, 1993.
2. Shiro Ise, “A principle of active control of sound based on the Kirchhoff-Helmholtz integral equation and the inverse system theory”, The Journal of Acoustical Society of Japan, Vol. 53, No. 9, pp. 706–713, 1997.
3. W. Hashimoto and H. Iwata, “Ensphered vision: Spherical immersive display using convex mirror”, Trans. of the Virtual Reality Society of Japan, Vol. 4, No. 3, pp. 479–486, 1999.
4. Atsunobu Kaminuma, Shiro Ise, and Kiyohiro Shikano, “Sound reproduction-system design considering head movement (in Japanese)”, Trans. of the Virtual Reality Society of Japan, Vol.5, No.3, pp.957-964, 2000.
5. K. Yamazawa, and H. Takemura, and N. Yokoya, “Telepresence system with an omnidirectional HD camera”, Proc. of Fifth Asian Conference on Computer Vision(ACCV2002), Vol. II, pp. 533–538, 2002.
6. R. Swaminathan, S. K. Nayar, and M. D. Grossberg, “Designing of Mirrors for catadioptric systems that minimize image error”, Proc. of IEEE Workshop on Omnidirectional Vision (OMNIVIS), 2004.
7. S. Ikeda, T. Sato, M. Kanbara, and N. Yokoya, “Immersive telepresence system with a locomotion interface using high-resolution omnidirectional videos”, Proc. of IAPR Conf. on Machine Vision Applications(MVA), pp. 602–605, 2005.
8. S. Enomoto, Y. Ikeda, S. Ise, and S. Nakamura, “Three-dimensional sound field reproduction and recording system based on the boundary surface control principle”, The 14th Int. Conf. on Auditory Display, pp. o_16, 2008.
9. H. Azuma, Y. Mukaigawa, and Y. Yagi, “Spatio-Temporal Lifelog Using a Wearable Compound Omnidirectional Sensor”, Proc. of the Eighth Workshop on Omnidirectional Vision, Camera Networks and Non-classical Cameras (ONIVIS2008), 2008.
10. Kazuaki Kondo, Yasuhiro Mukaigawa, and Yasushi Yagi, “Wearable Imaging System for Capturing Omnidirectional Movies from a First-person Perspective”, Proc. of The 16th ACM Symposium on Virtual Reality Software and Technology(VRST2009), 2009.